FLOSSmole, FLOSShub and the SRDA Repositories

Past, Present, and Future

Greg Madey
University of Notre Dame

Megan Squire
Elon University

FLOSS Community Metrics Meeting
Portland, Oregon, July 20, 2014

SRDA

FLOSShub

FLOSSmole

Who do we serve?

- Mostly academics: graduate students, faculty, post-docs, class projects, etc.
- Researchers interested in: FLOSS, software engineering, sociotechnical processes, social networks, data/text mining, economics, management information systems, open processes, citizen science/ engineering, innovation, …

How do we serve?

- Collect, archive, curate meta-data about FLOSS: project and developer statistics, discussion forums, bug/issues, releases, developer roles, project governance, project evolution, success criteria, history, …
- Mostly data from Forges …

We started with Sourceforge data

1999

2004

2007

2014

SRDA & FLOSSmole

- SRDA started in 2002
  crawler, direct data dump ⟹ *PostgreSQL*
- FLOSSmole started in 2004
  crawlers ⟹ *MySQL*

Index of /data/sf

Name	Last modified	Size	Description
Parent Directory		-	
2004/	02-Oct-2013 11:58	-	
2005/	02-Oct-2013 13:18	-	
2006/	02-Oct-2013 11:57	-	
2007/	02-Oct-2013 11:58	-	
2008/	02-Oct-2013 12:30	-	
2009/	19-Sep-2013 10:13	-	

FLOSSmole/SourceForge data

- January 2003 - sf0103
- November 2004 - sf1104
- December 2004 - sf1204
- February 2005 - sf0205
- March 2005 - sf0305
- April 2005 - sf0405
- May 2005 - sf0505
- June 2005 - sf0605
- July 2005 - sf0705
- August 2005 - sf0805

…

- August 2013 - sf0813
- September 2013 - sf0913
- October 2013 - sf1013
- November 2013 - sf1113
- December 2013 - sf1213
- January 2014 - sf0114
- February 2014 - sf0214
- March 2014 - sf0314
- May 2014 - sf0514
- June 2014 - sf0614

SRDA
112 dumps

2002-2009

- Focus was on collecting metadata, storing it

  SRDA collects from SourceForge

  FLOSSmole began collecting from other forges as well
  - Freshmeat
  - Rubyforge
  - Objectweb
  - Free Software Foundation
  - Alioth
  - Launchpad
  - Tigris
  - Google Code
  - Github

2002-2009

- This period was "The Age of the Forge"

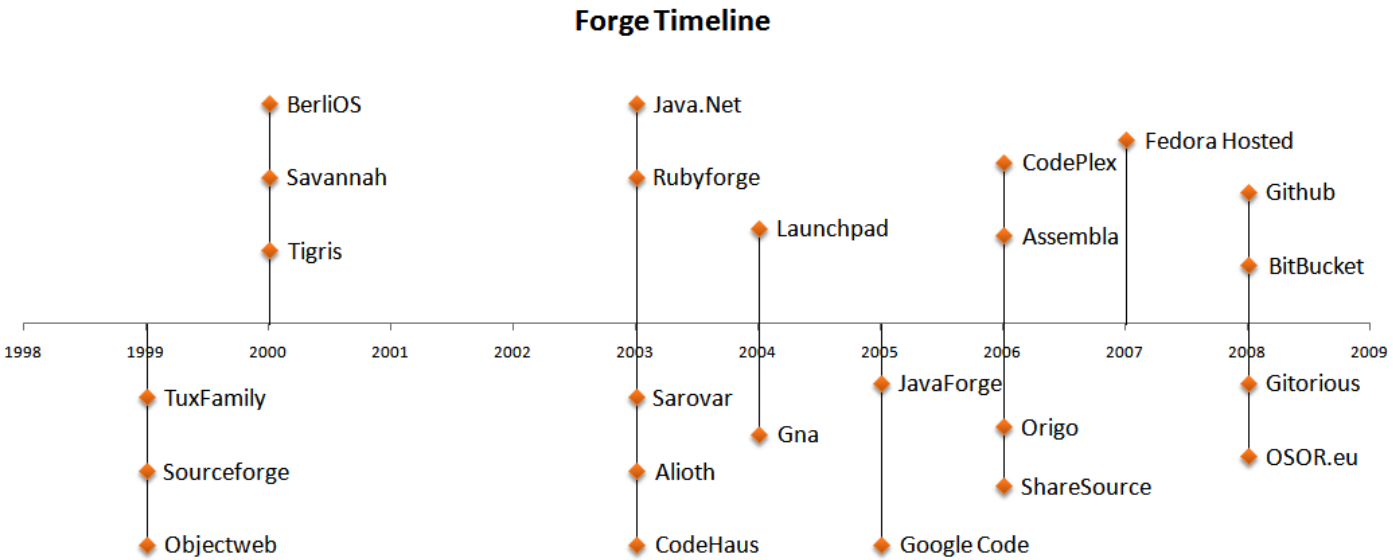
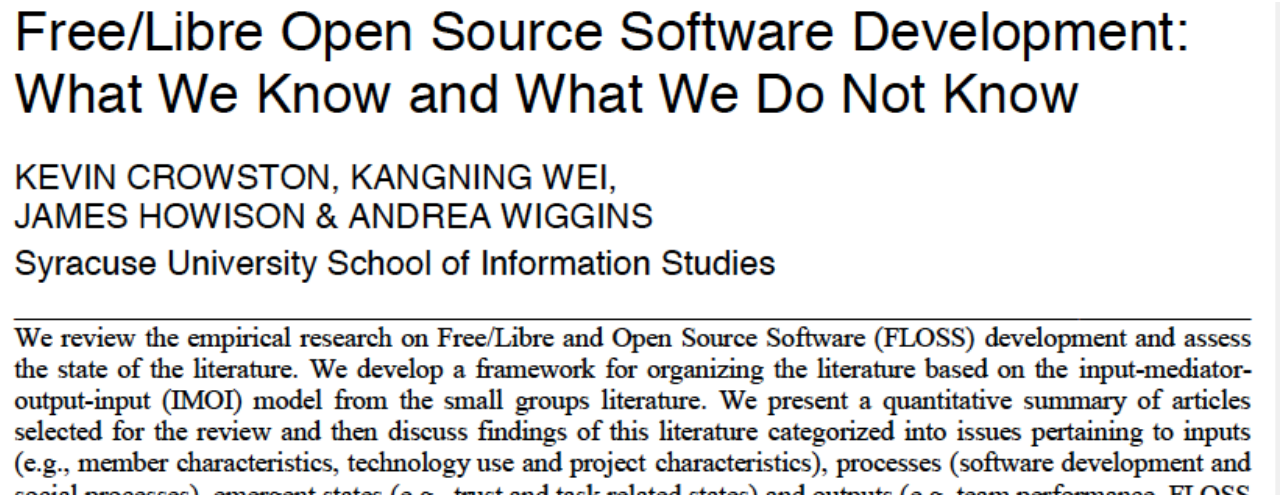


image from [Squire & Williams, 2012](#) (PDF)

2002-2009

- Growing our user base & supporting our users so they can conduct research using our data and write scholarly papers
- Example:

Crowston, Kevin, Kangning Wei, James Howison, and Andrea Wiggins. "Free/Libre open-source software development: What we know and what we do not know." *ACM Computing Surveys (CSUR)* 44, no. 2 (2012): 7



Free/Libre Open Source Software Development: What We Know and What We Do Not Know

KEVIN CROWSTON, KANGNING WEI,
JAMES HOWISON & ANDREA WIGGINS
Syracuse University School of Information Studies

We review the empirical research on Free/Libre and Open Source Software (FLOSS) development and assess the state of the literature. We develop a framework for organizing the literature based on the input-mediator-output-input (IMOI) model from the small groups literature. We present a quantitative summary of articles selected for the review and then discuss findings of this literature categorized into issues pertaining to inputs (e.g., member characteristics, technology use and project characteristics), processes (software development and

2002-2009

- We finally all met in the same room at the same time! (Limerick OSS 2007)
- 2nd WoPDaSD, 2007 (organized by Jesús González-Barahona, Megan Squire, Gregorio Robles)
- Kevin Crowston
- Greg Madey
- Walt Scacchi

2002-2009

Both teams won NSF grants to support work

- ISS-0222829
- CNS-0751120
- CNS-0708437
- CNS-0708767

these were Greg's

these were Megan's & Kevin Crowston's

In 2009-10 a few things happened...

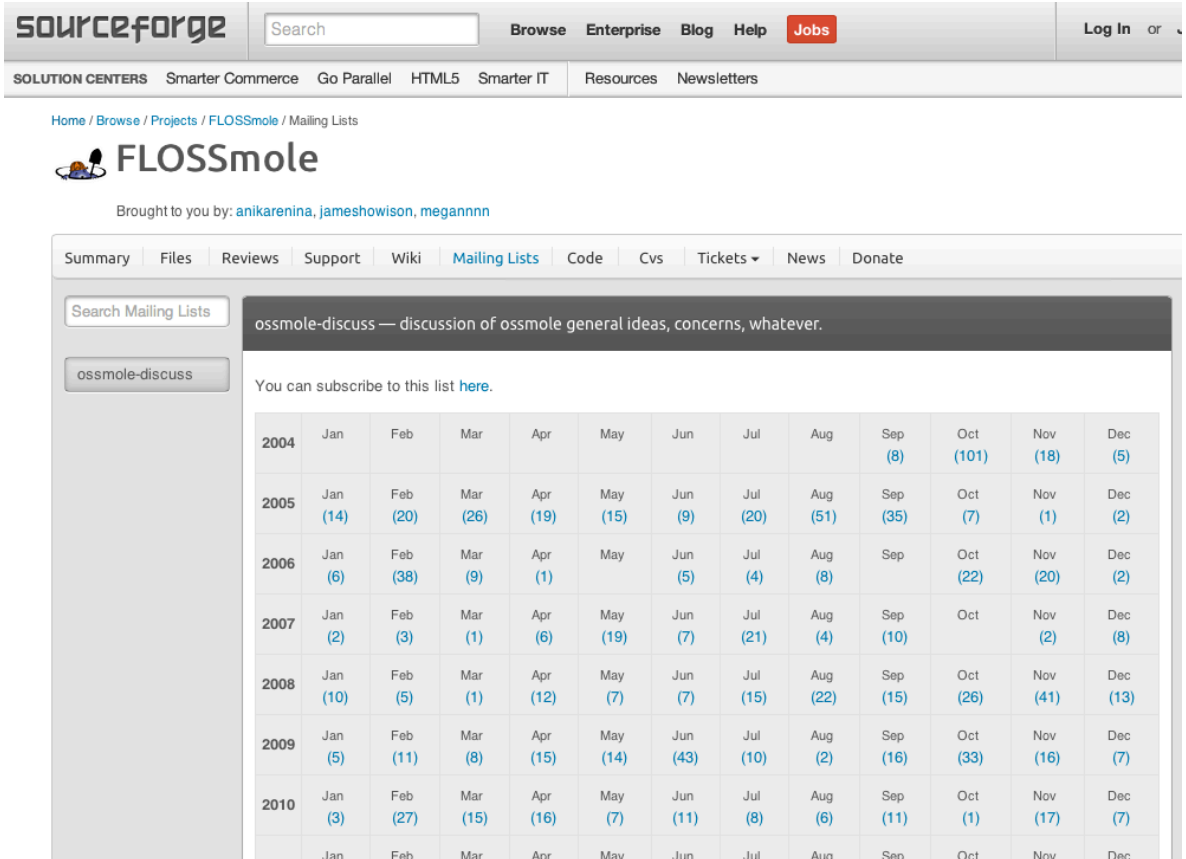
1. FLOSSmole stopped collecting from SF

   Due to a major SF site overhaul that broke the crawlers - again

In 2009-10 a few things happened...

2. SRDA & FLOSSmole merged mailing lists

We acknowledged that our users are largely the same



In 2009-10 a few things happened...

3. We started working on CRA-CCC grant with Walt Scacchi

   This made us more of an intentional community

Scacchi, Walt, K. Jensen Crowston, Chris Jensen, Greg Madey, Megan Squire, Thomas Alspaugh, Les Gasser et al. "Towards a science of open source systems." (2010).

In 2009-10 a few things happened ...

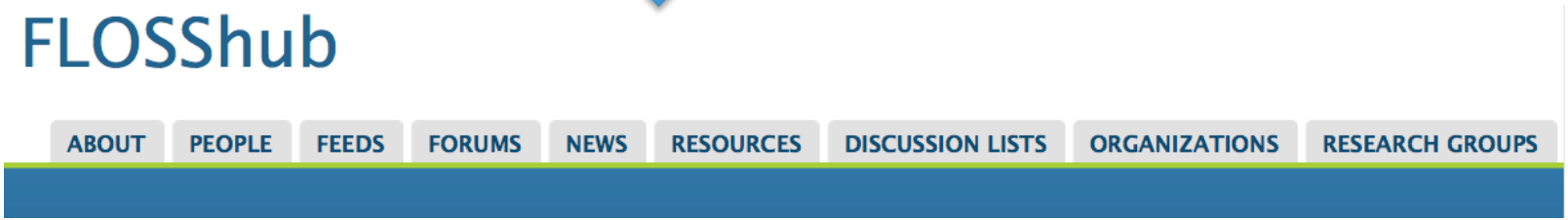
4. OSS 2010 conference was held in US at University of Notre Dame

Several meetings and collaborative efforts, joint presentations, panels, etc

In 2009-10 a few things happened...

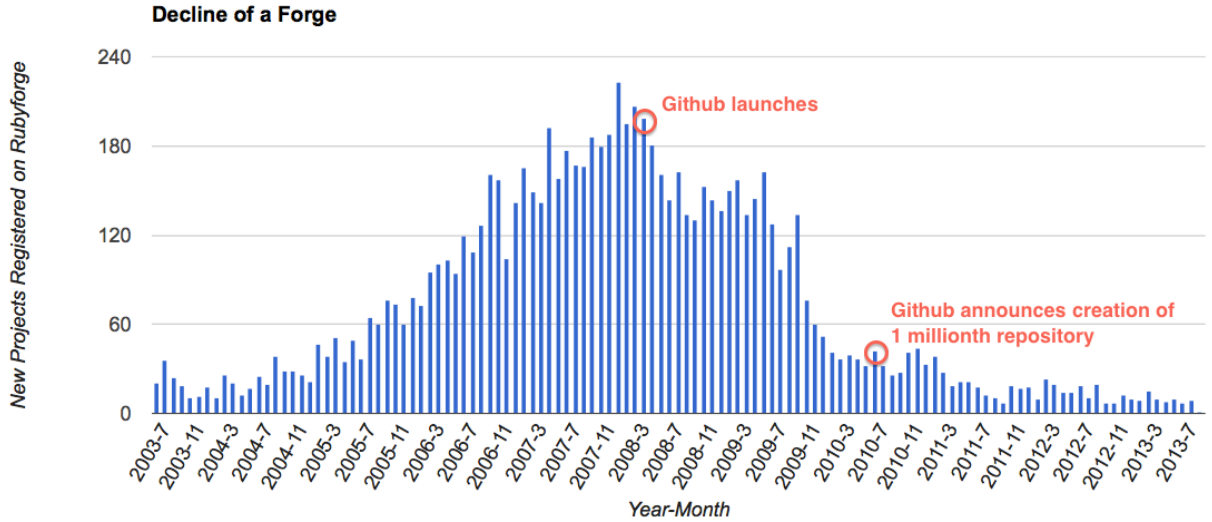
5. We created FLOSShub/biblio

- *We inherited the MIT F/OSS paper repository (Lakhani, von Hippel, and Hill)*
- *Since then we have added about 1000 papers to it*



In 2009-10 a few things happened...

6. The ascendance of Github and the decline of other forges begins



Graph shows the decline in new project registrations on Rubyforge

more info on this image

Statistics …

SRDA

- ~150 paper cites
- ~470 registered user accounts (perhaps double the number of users)

FLOSSmole

- ~200 paper cites
- unknown # of users
  - 203 on mailing list
  - 43 *active* db user accounts
  - hundreds of thousands of downloads

Current Machines & Sizes

SRDA:

- Currently on 2nd gen machine: 2 cores, 12GB memory, 7TB storage, SRDA data = 1.5TB
- New machine: 64 cores, 128GB memory, 12 TB storage

Current Machines & Sizes

FLOSSmole:

- Currently several terabytes of data stored at Elon (development server) and Syracuse (production server)
- We briefly hosted at San Diego Supercomputer Center and on Amazon Web Services

More recently...

FLOSSmole began shifting gears

- Away from metadata and away from forges
  due to the ascendance of Github & newer tools (Forge++ paper)
- Towards more complicated artifacts
  collecting & analyzing text, especially developer communication: email, IRC, Stack Overflow, Twitter
- Primary focus is still on building re-usable data sets

Goals for 2014 & beyond

1. Continue integrating SRDA & FLOSSmole
   Grant funding (supposedly forthcoming) will help
2. Continue FLOSS focus
   FLOSS is no longer a "weird" phenomenon, but it's still worth studying
3. Continue data focus
   - *Our primary job is data collection, archiving, and curation*
   - *Our secondary job is analysis*

Examples of current focus areas

Data source: Email
Project: Apache, LKML

Example 1:
Review of how 72 FLOSS research papers have used email archives
Example 2: Using email archives to study developer reactions to an innovation (paper under review)

Examples of current focus areas

Data source: IRC logs
Projects: Ubuntu, Apache projects, Django, etc

Example: Using FLOSS developer dialogue to train natural language classifiers to discern humor, insults, and profanity (under review)

All IRC data is available on FLOSSmole; some as flat files and all in MySQL

Examples of current focus areas

Data source: Messy metadata
Project: Apache

Example: Use the
Apache Board meeting minutes (very messy!)
to create data sets of who's-who on Apache
(paper link & data set)

Examples of current focus areas

Data source: Twitter
Projects: all FLOSS

Example: In order to mine Twitter, we need to know WHO are the FLOSS developers/projects on there (paper link & data set)

Examples of newer focus areas

Data source: Stack Overflow

- *Critically important as a "Forge 2.0" code/knowledge repository*
- *Not specific to FLOSS, but very important to both FLOSS and non-FLOSS*

Example 1: MSR challenge 2013 was on Stack Overflow

Example 2:
"A Bit of Code": How the Stack Overflow community creates quality postings

Example 3: Many FLOSS projects are outsourcing their developer support to Stack Overflow. Does this work? (paper under review)

Growth Areas

- ***Text sources*** - FLOSSmole is particularly interested in social coding and communication artifacts

- ***Donations*** - FLOSSmole is starting to get more donations of data, although there is still a lot of resistance

Ongoing Challenges

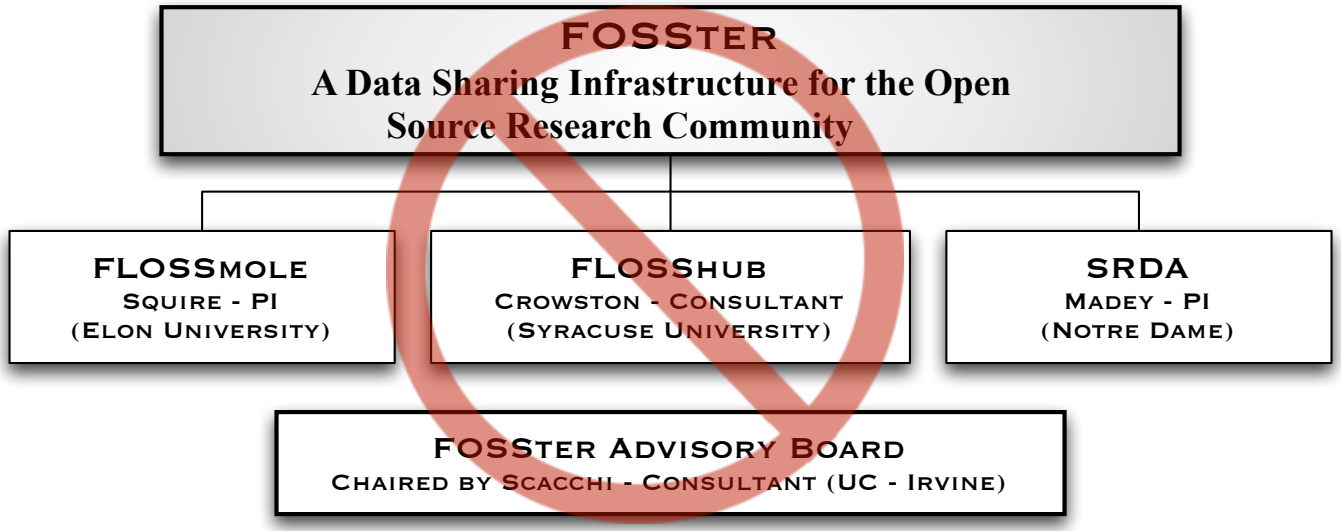
- ***Hardware*** - Constantly growing, need funding for machines, backups, disks

- ***Software*** - New data sources don't always fit in relational model; newer analysis methods need more sophisticated software

- ***Access & Support*** - community always needs more/ better support, docs, examples, PR, **integration**

Ongoing Challenges

- ***Funding to enable integration … several fails, tentative small success!***



Thank You!

Questions?

http://flosshub.org
http://flossmole.org
http://srda.cse.nd.edu/

SRDA

FLOSShub

FLOSSmole